情报检索中的文字辨析

陈　颖

（河北大学）

1990.6.25

文字辨析理论的基本要点

文字辨析是处理和查阅外文图书情报、文献资料的第一个环节，是外文图书情报工作者和查阅多种文字资料的科技工作者所必须具备的业务能力。

文字辨析的目的就是要通过对各种不同文字系统的对比分析，找出某种文字与其它文字不同的显著特点，从而判定出这种文字所对应的语种，达到辨识某种文字所属语言的目的。从这种意义上来说文字辨析理论是对比语言学理论的一个组成部分。另外，从文字学的角度出发，我们甚至可以把文字辨析的理论系统称做为对比文字学。因此，可以说文字辨析是文字学和对比语言学两个学科交叉点上的一个小学科。

这里我们还要论及语文对比中最基本的概念。

为了方便起见，我们把需要进行对比的语言（指语声所形成的口头语言）或需要进行辨识的文字（实际上是文字所组成的书面语言）统称语文。也就是说在这里语文是语言文字的简称，是书面语言的别称。进而，我们还可以把对比语言学或文字辨析都看成是语文的对比。无论在对个别模式的对比，还是对“系统”的对比中凡是构成一组“对比”，至少需要有两项，不妨叫做对比项。在这两个相互对比的对比项中，做为参照标准的那种语文叫做基准语文，而被比较的那种语文则叫做目的语文。

既然语文是语言文字的统称，文字辨析是通过分析记录语言的文字来判定文字系统所对应的语言系统，那么在文字辨析的过程中，既会牵涉到文字的类型和谱系，也会牵涉到语言的类型和谱系。因

而。在探讨语文的谱系分类这个课题之前，必须明确指出，所谓语文的谱系并不是指语言与文字组合体的共有谱系，而是分别指语言谱系与文字谱系，是这两个谱系的统称。这里还有必要将语言与文字的类型和谱系做一个概括的说明。

世界上有数千种（三、四千种，还是七、八千种，其说不一）语言，然而世界上百分之九十五以上的人所讲的语言也只有一百种左右。至于记录这些语言的文字却没有和语言呈一一对应的状态。这也就是说：第一，有些语言有文字，有些语言没有文字，而且至今绝大多数语言也还没有文字，只有一百余种语言有正式的文字；第二，同一种类型（甚至同一种）的语言对应着不同类型的文字，而同一种类型的文字又对应着不同类型的语言。

世界上的语言，按其结构从形态学上分成孤立语、粘着语、屈折语和编插语几种主要类型。而语言的谱系分类则是，以语言的上述类型为依据，用历史比较法找出诸语言间的亲缘关系编排成语言谱系。在语言谱系分类中，由大到小的各级单位分别是语系、语族、语支和语种。现在世界上约有二十个主要的语系和五十多个次要语系。其中使用人数最多的语系是：汉藏语系（孤立语类型）、印欧语系（屈折语类型）、含闪语系（屈折语类型）、阿尔泰语系（粘着语类型）。

世界上的文字，因为所取标准之一，对其类型的划分目前尚有争论。传统上一般把世界的文字分为象形文字、表意文字和表音文字（包括音节文字和音位文字）三种类型；近年来，有的学者提出，以书写符号的意义为根据，把世界上的文字分为句意字、表词字、词素字、音节字和音位字（又称音素字，包括辅音音素字和元音音

素字）五种类型。然而，无论以哪种类型体系为依据，对文字进行谱系分类，其结果都一样。现在世界上主要有四大文字谱系，此外还有十种不属于这四个文字谱系的文字体系。四大文字谱系是：拉丁文字系统（元音音素字母）、斯拉夫文字系统（元音音素字母）阿拉伯文字系统（辅音音素字母）、印地天城体文字系统（音节字母）。另外的十种文字是：希腊文、格鲁夫亚文、亚美尼亚文、埃塞俄比亚文、叙利亚文、希伯来文、蒙古文、朝鲜文、日本文、汉字。还有几种残存于非洲、亚洲和美洲印安人中的文字处于消亡状态。

为了识别使用同一种文字的不同种语言，我们把目的语文分成两类：被识语文和被辨语文。

第一类是识别不归属于当代世界四大文字谱系的那十种文字，即汉文、日文、朝鲜文、埃塞俄比亚文、希腊文、亚美尼亚文、希伯来文、蒙古文、叙利亚文。对于这十种语文中的任何一种，只要参照世界各种文字的文字样品，就能按照其行文的字形予以识别。这些语文没有辨析意义，只有识别意义，我们把这些只需要识别的目的语文称为被识语文或被识文字。

第二类是识别可归属于当代世界四大文字谱系——拉丁文、斯拉夫文、阿拉伯文、印地文——的某种文字。对于这四大文字谱系中的任何一种文字，我们首先要对照世界各种文字样品，初步确定其归属于哪一个文字谱系；然后再以所属谱系中的某种语文做基原语文（通常选择通用的或自己熟悉的），对照比较，找出目的语文的特征，进行辨析，才能识别出目的语文到底是记录了语言谱系中的哪一种语言。这些语文不但有辨析意义，而且也有识别意义。

我们把这种需要经过辨析才能识别的语文称为被辨语文或被辨文字。当然，广义地说，被辨文字也属于被识文字。文字辨析的探讨重点是第二类被辨文字。

我们以对比理论基础中的可比性为前题，得知在第二类被辨语文中，只有属于同一文字谱系的目的语文才能进行对比，而属于不同谱系的目的语文，或不属于四大文字谱系的目的语文就不能进行对比。在同一种谱系中，我们所选的基原语文往往需要具有如下条件①文字规范：尤其要求其字母表中带有的变体字母或附加符号尽量少些。②使用悠久：即这种语文被使用的历史长一些这好。③分析广泛：即使用这种语文的国家（或地区）和人数多一些为好。④掌握熟练：对辨析者来说，要尽量选择本人熟悉的语文。由这四点可见，在拉丁文字谱系中，我们常选英文为基原语文；在斯拉夫文字谱系中，常选俄文；在阿拉伯文字谱系中，常选阿拉伯文；在印地文字谱系中，常选印地文。

在进行语文的具体辨析时，我们要注意到文字谱系群与语言谱系群是否相匹配的问题，尤其在当代世界上许多文字都通过泛拉丁化趋向世界型、国际化的潮流中，拉丁文字母已经成为六十多种语言所使用的文字，这一点更值得关注。

附一将四大文字谱系的字母表附于此后，以供参阅。

在对具体的语文进行辨析时，我们首先要对被辨析的目的语文进行行文分析，以便使目的语文和基原语文的各个层次相互对应，只有在这两种语文的同一级单位上，按照可比性前题，才能加以对比。

要想准确无误地确定被辨语文在语言谱系分类中所记录的语种，

对被辨语文做如下的行文的层次分析，是一个行之有效的方法。

单字

因为语言的基本单位，在口头语言体系中是组成语句的单词或词素，与此相应，在记录口语的书面语言体系中，其基本单位就是组成字句的单字。所以说，单字是文字体系中行文分析的基本单位。每个单字可以记录一个单词，也可以记录一个词素；每个单字可以是单音节，也可以是多音节。

字母

在音节文字中，音节由音节字母来表示；在音素／音位文字中，音素／音位也由字母来表示。在这两种文字中，由音节组成的单词，或由音素组成了音节之后再组成的单词，都是用字母所组成的单字来表示，进而再由字母组成的单字来组成记录语句的字句。音节和音素／音位是语言的最小单位，字母是记录语言最小单位(音节和音素／音位)的书面形式，那么从文字辨析的角度来看，字母就应该是文字体系中行文分析的最小单位，即下限单位。当然每个字母都有变体，字母表每个字(母)位(置)上都包括这种字母的不同字体的大写和小写的字母变体形式。为了与语音分析对应，我们不妨认为字母的另一个名称为字素／字位。

字形特征

不同的字母之所以不同，是因为不同字母之间有区别功能，既然字母具有区别功能，那么任何一个字母就必然有自身独立存在的、不同于其它字母的区别特征，否则就不会成为不同的字母，由此可见，行文分析不能停止在字母这个层次上，我们还要更进一步把字

母分析为具有区别特征的最小区别单位。我们把构成字母的最小区别单元称为字母书写（即行文）时的字形（结构、笔画）特征。任何一个字母都是由一组字形特征所组成的。每个不同字母之间至少有一个字形特征不同。

对不同的文字谱系来说，其字形特征中笔画形状的处理方式不同。汉字是由横平竖直的直形笔画组成的方块字。拉丁文和斯拉夫文是由平滑连续的弧线笔画组成。阿拉伯文笔画形态是曲折多变。点逗斑驳。印地天城体文字的最富有特征的笔画是每个弯析线字母组成的单词上方都有一条连续的横线把这些字母连在一起。

超字形特征

在实际的行文过程中，对各个层次，不只需要上述的静态分析，而且也要进行动态分析。这样我们就要引入超字形特征这个概念。行文的超字形特征包括笔顺规则，书写顺序，大写规定，移行规则，字母的拼写附加符号，行文的标点符号等。这些行文过程中的超出字形特征之外的超字形特征。

标识项目

这样在行文分析中，就像在语音分析、语义分析和语法分析中一样，我们在行文的上限单位——字句，基本单位——单字，下限单位——字母——最小单位，最小区别单元——字形特征以及超字形特征，这些层面上对某种文字进行行文分析，以便系统地进行文字辨析。继而，在这些层面上我们可以通过对比分析，找出下列显示某种文字特征的标识项目。

在对比字母时，我们可以找出某种语文的特殊字母（包括附加拼写符号的字母）做为标识字母和标识符号，找出这种语文中的常

用字母组合做为标识组合。在对比超字形特征时，可以找出某种标点符号的特殊使用形式做为标识标点。找出有关大写等等特殊规定做为标识格式。在对比单词时，可以找出这种语言的冠词、连词、介词等短小功能词做为标识单词。当然如果能掌握某种语言中最常用的关键词（除标识词外，还有相关词和序列词。相关词包括表示疑问和指示的代词和副词。序列词包括表示时间、空间顺序的付词和名词），对文字辨析和阅读均有所裨益。

下面列出文字辨析中不同层次的各个标识项目，以便对各个谱系的每种语言进行量与质的分析，达到文字辨析目的。

语种名称：按照语言分类谱系表

文种名称：按照文字分类谱系表

文字样品：

字母表：（说明每个字母名称和字母总数，元音、辅音数）

标识字母：（有多少与基原语文不相同的字母）

标识组合：（特有的常用字母组合）

标识单词：（包括功能词、相关词、序列词）

标识符号：（包括拼写符号和标点符号）

标识格式：（包括大写规定、书写顺序等）

考虑到拉丁字母系统是当今世界上最大的文字谱系，而且图书情报工作中，西文书目的著录卡片是按拉丁字母顺序将各个语种混合排在一起的，故附二以科技情报中使用最广泛的几种语言的文字为例，来说明表中的各个标识项目。

参　考　文　献

1、Georg F.von Ostermann & A.E Giegengack:

Manual of Foreign Languages.New Tork, 1985

2、吉利亚列夫斯基、格利夫宁编、杜松寿译

世界各种文字样品　　北京　1958

3、中西亮著、付全铎译:

世界的文字　　北京　1985

4、肯尼思·卡兹纳著、黄长著、林书武译

世界的语言　　北京　1980

5、陈颖

文字辨析理论初探　　河北大学学报1988增刊